Corpus Methods in Philosophy

Joe Ulatowski, Dan Weijers, and Justin Sytsma

Philosophers commonly make claims about words or the concepts they are taken to express. Often the focus is on "ordinary" words (or concepts), although philosophers have also been concerned with technical terms. Sometimes engagement with concepts is the purpose of the research, as when a philosopher offers a conceptual analysis. Sometimes it serves as background, with philosophers laying out a concept in order to argue that it should be revised. And sometimes it is more instrumental, with conceptual issues arising while philosophers pursue non-conceptual questions.

Given this, it is (perhaps) surprising and (decidedly) unfortunate that philosophers do not generally employ tools designed for the systematic observation of the use of words. This is slowly beginning to change, however, with philosophers increasingly making use of corpus methods. While we cannot hope to detail the range of corpora or corpus methods that have been employed by philosophers, let alone the range that could be employed, in this short post we will offer a brief overview of corpus methods in philosophy, focusing on our own work. Interested readers can find a range of examples in the video presentations recorded as part of the recent Corpus Fortnight event (https://www.axphi.org/corpus-week) and in a number of overviews focused on philosophical use of corpus methods (Bluhm 2016, Mejía-Ramos et al. 2019, Sytsma et al. 2019, Caton 2020).

As experimental philosophers, each of us has begun to integrate corpus methods into our own research, whether simply to get a "feel" for the use of a term or to more carefully test linguistic hypotheses. One advantage of corpus methods for experimental philosophy is that they can offer a further way to test our hypotheses, one free of some encumbrances common in more standard experimental contexts, even as it inevitably introduces others. While this is far from the only benefit philosophers can (and have) derived from the use of corpus methods, it is the one that we focus on here.

I.

Corpus linguistics is a sub-discipline of linguistics that aims to collect and analyse existing, "real world" linguistic data (Biber et al. 1998, McCarthy and O'Keefe 2010, McEnery and Wilson 2001). Corpus analysis involves corpora: collections of written or oral texts. Corpora are typically curated, aiming to give a balanced and representative picture of the target domain of language use. The domain might be relatively general or highly specific, focusing on only certain types of texts or utterances (e.g., articles from specific disciplines, texts from certain time periods, the utterances of children). In addition, corpora often include further information, such as the base form of the words, part of speech, or syntactic structure, as well as various types of metadata, such as the source of the text or the age of the person making the utterance.

Advances in computing and the advent of the Internet have enabled corpora to grow in size and number and to be made freely accessible around the world. One of the most commonly employed English-language corpora among philosophers is the Corpus of

Contemporary English (COCA), which is comprised of over 1 billion words. COCA, along with a number of other corpora, is available from https://www.english-corpora.org/. Sophisticated search tools have been developed for such corpora, allowing users to easily determine how often a word, lemma, or phrase occurs in the corpus, the contexts in which it occurs, and more.

A couple of brief illustrations are in order. In a recent paper, one of us ran a number of studies looking at a range of attributions, including responsibility attributions, and performed a few simple searches using the non-academic portions of COCA as background (Sytsma ms). For instance, while Talbert (2019, 2) asserts that "in everyday speech, one often hears references to people's 'moral responsibility'," Sytsma found that the use of "moral responsibility" is in fact rather uncommon (69 occurrences compared to 9,501 for "responsibility"). In a related paper, Sytsma et al. (2019) used COCA to look at collocates for the phrases "responsible for the" and "caused the." They used the search features to determine which nouns most frequently occurred after each of these phrases, finding that they have a decidedly negative tinge (e.g., "death," "accident," "crash"), as confirmed by independent raters. In addition to giving occurrence counts for simple or sophisticated searches, COCA also provides the context for those occurrences. Fischer and Sytsma (ms) have made use of this in a recent paper, using COCA to create a random sample of 500 sentences with the word "zombie" in it, which was then used to assess the relative occurrence frequency of different senses of this term.

Other tools go further than mere searches. For instance, philosophers have employed topic-modelling algorithms to extract abstract topics for a collection of documents (see, e.g., Allen and Murdock (forthcoming) for discussion with regard to history and philosophy of science and Weatherson (2020) for application to philosophy journals). Other work has used distributional semantic models that map terms onto a geometric space based on the context in which they occur across the corpus and such that the distance between the term offers a measure of similarity of meaning. Continuing with the previous example, Sytsma et al. (2019) used the LSAfun package in R (Günther et al. 2015) to query a large premade semantic space (EN_100k_lsa), finding that "cause" and "responsible" were relatively close together and, in line with the previous findings, that they tended to be close to terms with a clear negative connotation, such as "blame" and "fault." Semantic models can also be used to investigate how word meaning has changed over time. For example, Ulatowski (ms) uses the Macroscope (Li et al. 2019; https://macroscope.tech/) to look at the meaning of "truth" diachronically.

The full text for many corpora is also available, including COCA, allowing one to build one's own semantic spaces for specific purposes. For example, Sytsma et al. built a semantic space using the non-academic portions of COCA with the phrases "caused the" and "responsible for the" treated as individual terms. They found that they were extremely close together in the resulting space, suggesting that they have similar meanings as captured by the contexts in which they are used in the corpus.

In addition to the large number of both general and specialty corpora available online, the internet can be used both as a corpus and for building corpora. While there is disagreement about how suitable the internet is for use as a corpus, standard web searches can provide at least some evidence for linguistic hypotheses, especially in conjunction with other corpora (for a few philosophical examples, see Knobe and Prinz 2008, Reuter 2011, Sytsma and Reuter 2017). The use of the internet for building a corpus, by contrast, is relatively uncontroversial. And a number of philosophers have used the web to create specialty corpora, such as compiling texts from *Philosophy of Science* (Malaterre et al. 2019), the works read by Darwin (Murdock et al. 2017), the works of Nietzsche (Alfano and Cheong 2019), and the two main online encyclopaedias of philosophy (Sytsma et al. 2019), among others. To give but one more example, a number of researchers have used JSTOR's Data for Research (www.jstor.org/dfr) to perform various searches on academic journals. For instance, Andow (2015) compares intuition-talk between philosophy and non-philosophy journals and Mizrahi (forthcoming) looks at the use of "truth," "knowledge," and "understanding" to explore how scientific practitioners conceive of scientific progress.

II.

Why should philosophers use corpora? As noted above, corpora can provide philosophers with examples of the use of words "in the wild." Insofar as philosophers put forward hypotheses or make assertions that either concern or generate predictions about word use among some population, the claims can be empirically tested. And corpus methods provide one valuable way of doing so.

Not surprisingly, as experimental philosophers we firmly believe that empirical claims call for empirical support. While we adopt a broad conception of experimental philosophy (Sytsma and Livengood 2016, Sytsma 2017), much of the work that has been done concerns "intuitions" and tests linguistic or conceptual claims. Most frequently this has involved the use of questionnaires, often with participants reading a short case description that mirrors "traditional" philosophical thought experiments, and then answering some questions about that vignette, generally including a question about whether a concept of philosophical interest applies in this case.

There are many worries that one may have about such experimental work. In particular, one could raise doubts about the relevance of the judgments elicited, argue that such judgments (or the "intuitions" they might be taken to reflect) do not or should not play any role in philosophical analysis, or argue that while such judgments do play a role in philosophical analysis, the judgments of philosophers are to be preferred over the judgments of lay people. Alternatively, one could raise doubts about the results based on worries about the experimental approach. One might argue that the experimental context raises the spectre of experimental artifacts: that the phrasing and presentation of the materials impacts how participants respond, perhaps biasing the results.

The use of corpus analysis can help to mitigate against such potential problems, bringing to bear another channel of evidence on linguistic and conceptual claims, with the possibility of a consilience of evidence that would increase confidence in each method. The critical thing to note is that corpora by-and-large involve "real world" linguistic data—texts and utterances produced outside of any artificial experimental context. The data isn't generated via vignettes and questions devised by an experimenter who might attempt to shape responses in a particular way, biasing responses toward the investigator's own views. The use of corpus methods, thus, can be a valuable addition to the more common experimental methods. It can aid initial exploration and hypothesis generation and it can help confirm experimental results, testing concerns about the impact of experimental artifacts.

The use of corpus analysis in philosophy doesn't come without its own challenges, however. Investigators should be aware of potential issues and use a range of methods to mitigate against these worries. As always, researchers should get clear on the hypotheses they are testing and the tools they are using to test. Searches will vary depending on the corpus used and what you are searching for, raising issues for interpreting frequencies and highlighting the need for relevant comparisons. Words typically have multiple forms and are used in multiple ways, which can generate challenges when it comes to testing claims about a specific sense. One could target a specific word, lemma, or lexeme depending on the corpus. A lemma is a group of all inflectional forms related to one stem that belong to the same word class. We group together forms that have the same base and differ only with respect to grammar, such as for example the singular and plural forms of the same noun, the present and past tense of the same verb, the positive and superlative form of the same adjective. A lexeme, on the other hand, is a lemma with a particular meaning attached to it, which is necessary to distinguish different senses of polysemous words. Distinguishing lexemes is a difficult task, but one that is often central to drawing a philosophical conclusion. Similar issues arise in using more mathematical techniques. For instance, semantic spaces will vary notably based on a large number of decisions, including the corpus used, how it is pre-processed, the algorithm employed and the settings for its variables.

This, of course, just scratches the surface: corpus linguistics is a large and complicated field. That said, even a rank amateur can begin to get some benefit from corpus methods today. Every philosopher can begin to integrate a simple exploration on COCA, for example, when questions about the ordinary language or concepts arise. And expertise builds over time.

References

Alfano, M. and M. Cheong (2019). "Examining Moral Emotions in Nietzsche with the Semantic Web Exploration Tool: Nietzsche." *Journal of Nietzsche Studies*, 50(1): 1-10.

Allen, C. and J. Murdock (forthcoming). "LDA Topic Modeling: Contexts for the History & Philosophy of Science." http://philsci-archive.pitt.edu/17261/

Andow, J. (2015). "How 'Intuition' Exploded." *Metaphilosophy*, 46(2):189–212.

Biber, D., S. Conrad, and R. Reppen (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bluhm, R. (2016). "Corpus Analysis in Philosophy." In *Evidence, Experiment, and Argument in Linguistics and the Philosophy of Language*, edited by Martin Hinton, 91–109. New York: Peter Lang.

Caton, J. (2020). “Using Linguistic Corpora as a Philosophical Tool.” *Metaphilosophy*, 51(1): 51-70.

Fischer, E. and J. Sytsma (ms). “Zombie Intuitions.” Presentation available at: https://www.axphi.org/corpus-week

Günther, F., C. Dudschig, and B. Kaup (2015). “LSAfun: An R package for computations based on Latent Semantic Analysis.” *Behavior Research Methods*, 47: 930–944.

Knobe, J., and J. Prinz (2008). “Intuitions about consciousness: Experimental studies.” *Phenomenology and the Cognitive Sciences*, 7: 67–85.

Li, Y., T. Engelthaler, C. Siew, and T. Hills (2019). “The Macroscope: A tool for examining the historical structure of language.” *Behavior Research Methods*, 51: 1864-1877.

Malaterre, C., J. Chartier, and D. Pulizzotto (2019). “What is this thing called *Philosophy of Science*? A computational topic-modeling perspective, 1934-2015.” *HOPOS*, 9: 215-249.

McCarthy, M. and A. O’Keeffe (2010). *The Routledge handbook of corpus linguistics*. London: Routledge.

McEnery, T. and A. Wilson (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mejía-Ramos, J., L. Alcock, K. Lew, P. Rago, C. Sangwin, and M. Inglis (2019). “Using Corpus Linguistics to Investigate Mathematical Explanation.” In E. Fischer and M. Curtis (Eds.), *Methodological Advances in Experimental Philosophy*, Bloomsbury, 239-264.

Mizrahi, M. (forthcoming). “Conceptions of scientific progress in scientific practice: An empirical study.” *Synthese*.

Murdock, J., C. Allen, and S. DeDeo (2017). “Exploration and Exploitation of Victorian Science in Darwin’s Reading Notebooks.” *Cognition*, 159: 117-126.

Reuter, K. (2011). “Distinguishing the appearance from the reality of pain.” *Journal of Consciousness Studies*, 18(9-10): 94-109.

Sytsma, J. (ms). “Crossed Wires: Blaming Artifacts for Bad Outcomes.” http://philsci-archive.pitt.edu/18293/

Sytsma, J. (2017). “Two Origin Stories for Experimental Philosophy.” *teorema*, 36(3): 23-43.

Sytsma, J. and J. Livengood (2015). *The Theory and Practice of Experimental Philosophy*. Broadview Press.

Sytsma, J., R. Bluhm, P. Willemsen, and K. Reuter (2019). "Causal Attributions and Corpus Analysis." In E. Fischer and M. Curtis (Eds.), *Methodological Advances in Experimental Philosophy*, Bloomsbury, 209-238.

Sytsma, J. and K. Reuter (2017). "Experimental Philosophy of Pain." *Journal of Indian Council of Philosophical Research*, 34(3): 611-628.

Talbert, M. (2019). "Moral Responsibility." *Stanford Encyclopedia of Philosophy*.

Ulatowski, J. (ms). "Does Truth Evolve? Diachronic Analysis from 1850 to 2010"

Weatherson, B. (2020). *A History of Philosophy Journals, Volume 1: Evidence from Topic Modeling, 1876-2013*. http://www-personal.umich.edu/~weath/lda/